Characteristics of the Happiest Countries in the World

Machine Learning with Gradient Descent and Clustering

Nelson Blickman
CS 613 Drexel University
Philadelphia, Pennsylvania nb939@drexel.edu

I. Abstract/Background

What makes a person happy? This is a question that keeps philosophers busy. There are some agreed upon general items such as good health, happy relations, and the feeling of self-worth, but every person is different, and there is generally not a steadfast rule that can answer the question. Moreover, what does it mean to be happy? Can it be calculated to a finite amount, say how many times in the month does your body possess a feeling of joyous energy, is it the quantifiable impact you have on your society? There is no clear answer to this either, however, the importance of this question is no secret of course. Without energy, optimism, high self-esteem, good character, happiness, life on earth would not move up or even forward. My introduction's purpose is not only to remind us of the grave importance of this topic which will pave the way for my algorithmic analysis, but also to tell the Philosophers to make some room, and let a data cruncher bring in some quantifiable calculations.

II. Related Work

The "World Happiness Report" is released once a year in accordance with the United Nations annual International Day of Happiness. It ranks the happiest countries in the world all the way down to the unhappiest. These reports allow the people of a country to report on how happy or content they are. Personal opinions of every day people guide the rankings, which of course opens the door for ambiguity. Adding to my introduction above, who is to say that they know whether or not they are happy, or happier than the next person? This question may be fair but personal opinions are not a bad place to start. From here, the World Happiness Report will collect data and understand the country itself. Is it wealthy with resources, are people free, do people perceive their neighbors as generous.

The Gallup World Poll and its researchers lead the effort in collecting the necessary data around the world that contributes to the content within the World Happiness Report. These reports have been going on for the past 8 years.

III. Analysis/Evaluation/Results

We can begin our analysis sufficiently without anymore philosophy, so I will not go in depth on the difference or similarity between a happy person and a happy country. Our main goal is to take our happiness ranking of every country, the 6 features of that particular country, and see if we can figure out if correlations exist. For example, is there a certain feature that determines a countries happiness level more than others?

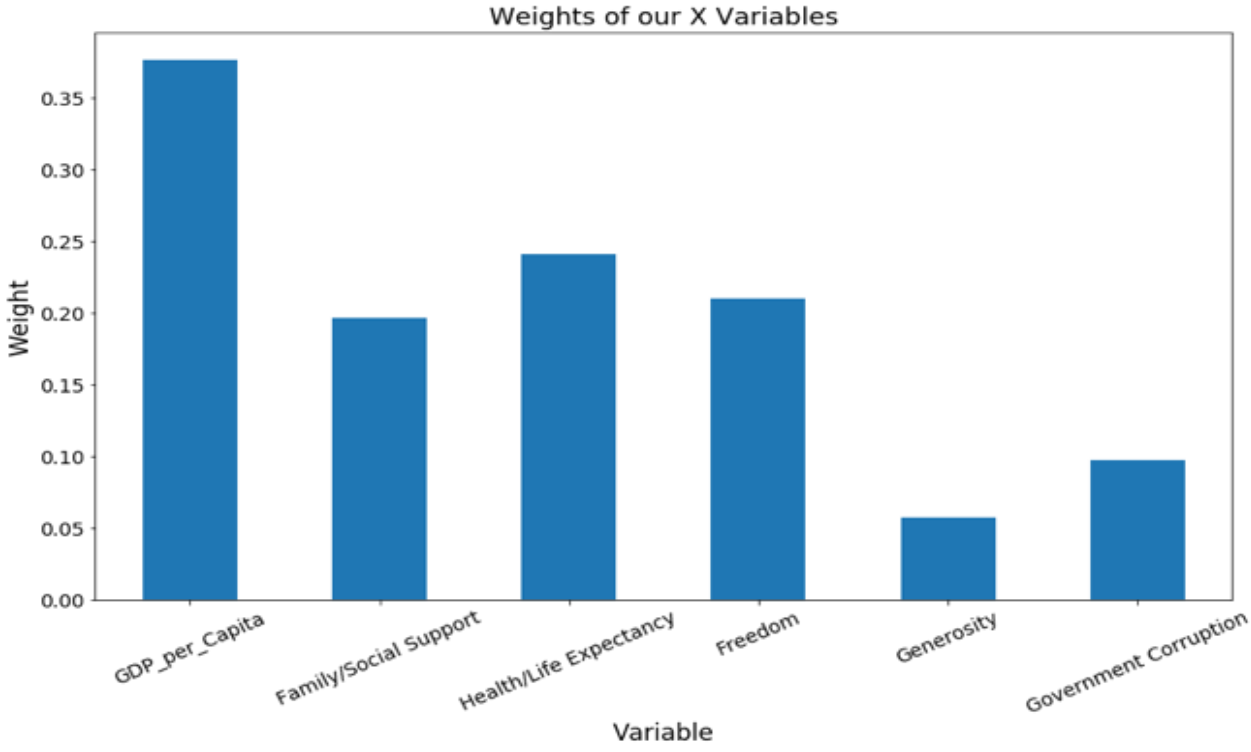
Data Cleaning

To begin, I imported, cleaned and organized data from the World Happiness Report. I imported this data directly from

Kaggle's website. The data consisted of the past 5 years (2015-2019) of rankings from 1 – 155 of a country. The independent features in the data varied in name from year to year (GDP vs GDP..per..Capita), which meant it required the use of regular expressions in Python in order compile all of the countries data over 5 years into one single Data Frame. From here we sliced the data into x and y variables, standardized, and split it. I have performed a few Machine Learning algorithms that I will discuss. Keep in mind that the potential Y variables that will be used may be the ranking of the countries overall happiness from 1-155, or the same ranking but with floating point numbers say 1.3 to 7.7 (the floating point ranking does not vary from ranking to ranking with equal length like the 1-155 ranking does by 1). Once again, these rankings are based off of the personal opinion of people living in that country. The potential X-Variables, our features, GDP per capita, health/life expectancy, freedom, government corruption, family/social support, and generosity, contain values that show how likely or unlikely it was to contribute to happiness of that country.

Gradient Descent

In our first Machine Learning analysis, we will use Gradient Descent in order to determine the optimal weights or significance for each of the 6 X Variables listed above. To begin, I use the linear regression cost equation, y=mx, with y as our floating point annual rankings, our six x-variables, and an initial set of random m's to find our error which in turn will give us our Root Mean Squared error. Next, we compute the dot product of our same x variables, but this time with our error, to give us the amount that we need to adjust our weights. After we update our weights, we repeat this process, march towards toward the minimum of our function until our error is incredibly low. It took almost 2,000 iterations to get to our optimal weights, our thetas, which can be seen below (ignoring the bias theta).

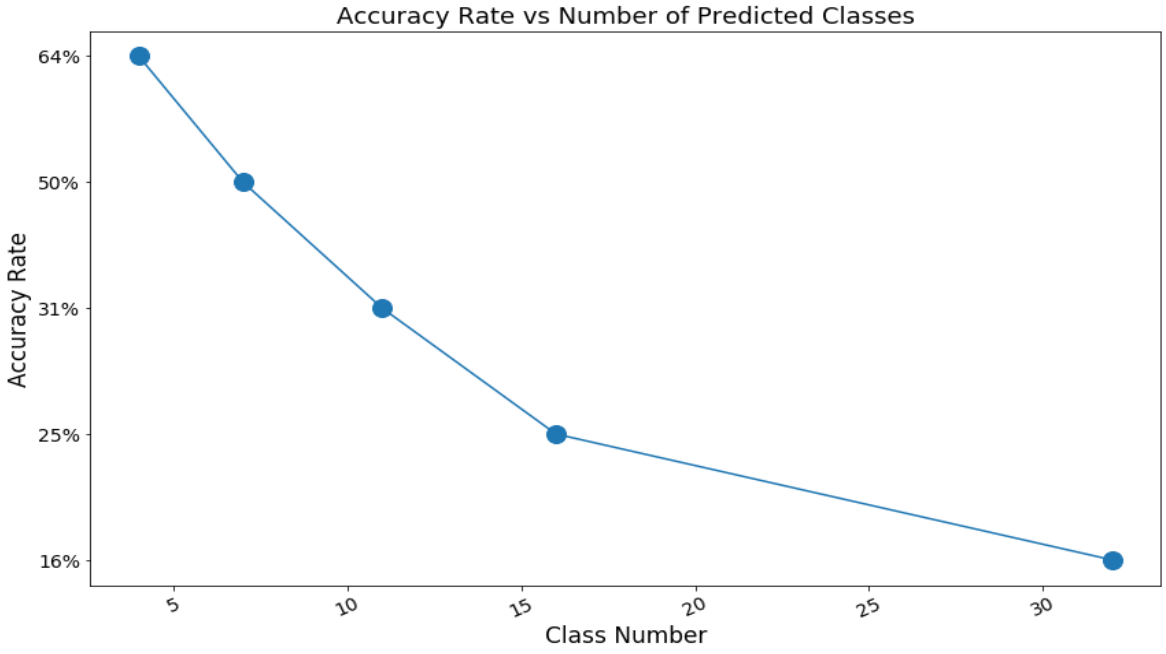


The results in the graph above give us insight. The optimal weights appear to show GDP, Health and Freedom as the most influential features. I don't think there is too much of a surprise here. The ability to live a healthy life, put food on the table, and have your freedom, are most likely to follow at the base of the famous "Maslow's hierarchy of needs" (For those unfamiliar, Abraham Maslow created a psychological theory in the mid 20th century that is still referred to today as a decent argument for the order in which humans need certain things). Generosity, and social support are important, but Maslow puts these items on a higher level (not as necessary) in the pyramid, so this is not a surprise either. I think Government Corruption seems a bit low. However, freedom may be an end goal of low government corruption, making it debatably more important.

K-Nearest-Neighbors

Next, we will use the K-Nearest-Neighbors (KNN) algorithm to understand our data, or what correlations exist within it. Our X-Variables are going to be the same 6 we used in our Gradient Descent analysis. However, our Y-Variable is now going to use the overall rankings from 1-155 (1, 2, 3… 155 with 1 being the happiest country). In addition, we first are going to group our Y variable ranking into 3 classes, and test the accuracy of the KNN algorithm as it must predict if a data point belongs to the first 52 happiest countries (1-52), in the 52-104 group, or in the 104-155 group. We will run this algorithm 5 more times, increasing the number of potential classes that our model could put our data point into, which means it would be getting increasing more difficult for the model to be accurate of course.

For every single testing data point, my KNN algorithm will compute the sum of squared difference from the current testing data point to every training data point. We will look at the y variable (happiness ranking 1-155) of the training data point that was the closest (since our k is equal to 1) distance to the testing point and the y variable of that training point is the models ranking prediction. If the prediction is ranking 47, the actual ranking is 32, and we are computing accuracy based on 3 different classes, than this was considered a correct prediction.



Our results in the above graph show us that when the model tries to predict if a sample data point belongs to 1 of 3 classes, it does very well. If we have to make predictions amongst a larger group of potential classes, the accuracy goes down. I think this is significant. It tells us that there are clearly correlations that exist in our data. We can predict the happiest countries, against the middle tier of happiest countries vs the unhappiest countries very well. It also shows us that predicting if a country should be in the top 5 or bottom 5, etc, in the rankings is not so easy. We may not be able to predict the exact rankings of a countries happiness.

Unsupervised K-Means Clustering

We will use the insight from our supervised KNN algorithm to help us create an unsupervised K-Means Clustering algorithm. Our KNN model was able to predict over 50% accuracy when it had to predict if a data point fell within 7 separate classes (of rankings). Therefore, in our K-Means analysis, we will initialize and create a total of 7 clusters. To do this, we begin by finding the distance between one data point and each of the 7 initialized k's (the random points of comparison for our 7 clusters). Next, the k with the shortest distance to the sample data point will acquire that data point as part of its cluster. After looping through every sample data point once, we will take the mean of each feature of each cluster, and update the value of the representative of the cluster to this mean value. We repeat this process until the change in our k representatives is very low.

After our algorithm runs, we should be left with 7 clusters, each containing a group of sample data points (countries with their happiness rank). From here, our code calculates and finds the 5 closest sample data points to the final k representative data point.

Cluster D:

Germany: Rank 15
Netherlands: Rank 6
Austria: Rank 13
Austria: Rank 12
Belgium: Rank 17

Average Happiness Ranking in Cluster: 12.6

(For a detailed review of the results for this analysis, please see Jupyter Notebook "final project code")

Shown above are the 5 closest sample data points (to k) of one of our 7 clusters, cluster D. The average happiness ranking shows 12.6. We may have expected this number to be 11.

With 7 clusters, out of 155 rankings, each would contain countries with rankings in groups of 22. Looking at Cluster D, the 5 closest sample data points would most likely represent the average rank of the first group of 22, which would theoretically be closer to the midpoint, thus 11.

Looking at the average ranking of all 7 clusters though, only a few of them did as well as Cluster D. The fact that our average rankings did not match what we would have expected for every cluster further shows us that there may not be a perfect correlations in our data. Granted this is still not bad. In our KNN analysis, we saw our accuracy shoot up when we grouped and predicted our country ranking for three classes. This idea that there is a strong correlation between the happiest countries, the average happiness countries, and the least happy countries is confirmed in our k-means clustering analysis. This is because two of our clusters (see code) have the exact same average ranking of 130 (of the 5 closest data points to the k representative of the cluster), two have the same average rank of 75, and two are very close to a rank of 15 (one of which was the 12.6 from Cluster D above).

These 6 clusters could have fit into 3 (1 of the 7 clusters not included).

IV. Conclusion

Overview of Algorithms: Based on our Gradient Descent Algorithm, our GDP, health/life expectancy, and freedom features have the highest weights. The KNN and Clustering models are telling us there may be the strongest correlation when we group the countries rankings into groups of 3. Therefore, GDP, health, and freedom are much higher or much lower amongst three groups of countries.

To elaborate on and validate this conclusion, we will calculate the mean and standard deviations of each of the three highly weighted features just discussed, and the other three lower weighted features (Family/Social Support, Generosity, and Government Corruption). But, we will do these mean/std calculations separately for the happiest 50 countries, middle 50 countries, and unhappiest 50 countries (more like 52 each to be exact). Then, I take the change in the mean of the values of those features when you go from one set of 50 countries to another set of 50 countries.

The numerical results of this are listed in the code. The results show that there is a larger difference in the means from cluster to cluster among our high versus low features. Since the difference in the means from group to group is larger in the high features, it shows that the high separation from each of the three clusters is due to those features. A country with high GDP in the first 50 happiest countries will have to have more than 1 economic recession take place to bring it down to the second 50 happy countries. (highlighting the fact that it s wise of us to put countries into one of three groups, based on the high weighted features)

Additionally, see the low standard deviations of the high (and low) features in my code. This also highlights the main point because it shows there is not a linear relationship between a countries happiness a nd its features. If you drop from the 10th happiest country to the 20th (and not the 50th) there may be no change in your features.

If you go down in GDP, health, or freedom, enough to another level, you may be a significantly less happy country. If you can increase these three features high enough, you will be significantly happier. This is not terribly surprising. However, it is insightful. It reminds us that there are certain things a country needs, for example a certain level of economic prosperity, to keep a countries people happy. And that, unfortunately, the unhappiest countries have a long way to go in terms of economic prosperity, health care, and individual freedom, in order to make the people of their country happy. There does not appear to be a short cut, like only focusing on policy that involves mutual cooperation or togetherness. Kindness, generosity, and support from friends and family can be powerful, but in a world filled with corruption, hunger, and sickness, we are reminded it is not enough.

IV. Future Work

The World Happiness Report already have plans to create rankings for 2021 and 2022, so clearly the analysis on the topic will continue. Many will continue to try and define what is it about certain countries that make the people living their happier than in other places. Of course, the bigger challenge is for those countries to actually make changes according to our analysis. It is easier to tell a country that if they can increase there GDP and life expectancy in their country than its people

will become happier, than it is for them to actually do it. However, the analysis is worthwhile not only because we may notice how we may be undervaluing the importance of having say good health care in society, but because it reminds us what is important in life, which can help foster action and ambition of the people to strive to obtain what they don't have.

References

"Find Open Datasets and Machine Learning Projects." *Kaggle*, www.kaggle.com/datasets?utm_medium=paid.

Gallup, Inc. "Tracking the World's Happiness." *Gallup.com*, Gallup, 2 Oct. 2020, www.gallup.com/analytics/247355/gallup-world-happiness-report.aspx.

World Happiness Report 2020, 20 Mar. 2020, worldhappiness.report/ed/2020/.